ISSN 1561-5359. Штучний інтелект, 2018, №3

УДК 004.853

В.М. Терещенко, А.Д. Бугайов
Київський національний університет імені Тараса Шевченка, Україна
пр-т. Глушкова 4д, м. Київ, 03680

АЛГОРИТМИ МАШИННОГО НАВЧАННЯ У КОНТЕКСТІ ВЕЛИКИХ ДАНИХ

V.M. Tereshchenko, A.D. Bugaiov
Taras Shevchenko National University of Kyiv, Ukraine
4d, Hlushkov ave, Kyiv, 03680

MACHINE LEARNING ALGORITHMS IN BIG DATA CONTEXT

Великі Дані обіцяють змінити наш звичний уклад повсякденного життя, роботи, відпочинку. Однак, вилучення інформації з великих масивів даних процес нетривіальний і досить ресурсомісткий. До того ж використовувати інструменти для аналізу даних, які були актуальні ще 10 років тому в сучасному контексті досить складно. У даній роботі розглянуті сучасні методи Машинного Навчання, які підходять для обробки Великих Даних, наведені їх переваги в конкретному середовищі і то як вони долають той чи інший виклик породжений Великими Даними. Врешті обрана одна методологія, яка досить широко покриває оголошені виклики і на ній зроблено акцент з коротким описом її проблематики в сучасному стані.
Ключові слова: Великі Дані, Машинне Навчання, Глибинне Навчання, Безперервне Навчання

Big Data promises to change our habitual way of daily life, work, leisure. However, extracting information from huge data sets is not a trivial process and is rather resource intensive. Furthermore, the tools for data analysis that were relevant 10 years ago aren’t so effective in the current context. In this paper is considered modern and popular methods of machine learning that are suitable for processing Big Data, addressed their advantages in a particular environment and described how they cope with the challenges coming from the Big Data. In the final analysis the methodology that broadly covers the announced calls was chosen and its current problems were described.
Keywords: Big Data, Machine Learning, Deep Learning, Lifelong Learning

Вступ

«Дані – це нове паливо». Так сказав Clive Humby, математик, архітектор Клубної карти Tesco з Великобританії у 2006. За останні більш ніж 10 років висловлювання лише набирає актуальності і повторюється у різних контекстах, різними людьми за різних обставин, але з тим же незмінним змістом. Причиною цього є постійно прибуваючі нові і нові дані. Дані несуть у собі інформацію. Крім відкритої інформації, тієї, що дані несуть користувачеві явно, існує прихована інформація, така як тренди, вподобання користувача, задоволеність товаром чи послугою і т.д., доступна тільки після ретельного аналізу величезних масивів цих даних. Кількість таких даних і їх джерел росте з кожним роком із запаморочливою швидкістю, даючи можливість аналізувати і виявляти приховану інформацію. Статистика з Інтернету, зокрема соціальні мережеві сервіси (SNS) свідчить, що кожну хвилину в Facebook публікується понад півмільйона коментарів, оновлюється близько 300000 статусів, і завантажується більш ніж 100000 фотографій [1]. Загальна кількість «твітів» за день у Твіттері – 500000000 [2]. Користувачі Інстаграм за день «лайкають» 4200000000 та діляться 95000000 постами [3]. Генерується незліченна кількість контентів з мобільних пристроїв і не тільки для SNS. Зростаюча тенденція на «підключені пристрої» Інтернету (IoT) і поєднане з нею генерування немислимої кількості нових даних. Для прикладу, у 2016 кількість «підключених пристроїв» була 6500000000, а до 2020 ця цифра обіцяє бути більш ніж 20000000000 [4] та кількість генерованої цими пристроями інформації на рік прогнозується перевищити 500 зетабайт (1 зетабайт = трильйон гігабайт). Не варто забувати і про інші домени, такі як медицина, роботи, сучасні автомобілі, дрони і т.д.

Одним із визнаних успішних методів обробки даних для отримання з них корисної інформації є Машинне Навчання (Machine Learning – ML). Існує твердження, що алгоритми ML вчаться витягувати інформацію із даних тим краще, чим більше даних для них доступно [5]. Однак, незважаючи на те, що потенційна вигода від використання Великих Даних значна і вже є реальні успіхи у даній області [6], як і раніше, залишається чимало технічних викликів та складнощів. Складнощі переважно зосереджені у тому, що традиційні алгоритми машинного навчання, такі як Метод Опорних Векторів (Support Vector Machine), Випадковий Ліс (Random Forest) та інші, створювалися без урахування подальшого їх використання в середовищі Великих Даних і були просто до цього не готові. Наприклад, безліч алгоритмів, що навчаються, розраховують на дані, які будуть повністю завантажені в пам'ять, для інших важливо, щоб тренувальний набір даних був повністю доступний на етапі тренування моделі [7]. У контексті великих даних це навряд чи можливо.

На етапі обробки великих даних і отриманні з них корисної інформації труднощі зустрічаються на кожному кроці. При отриманні та очищенні даних, при добуванні інформації, при інтеграції та поданні інформації, при моделюванні і аналізі, при інтерпретації результату тощо [6]. Для розуміння джерел складнощів у машинному навчанні на Великих Даних пропонується розглянути наступний набір проблем, що мають найменування 4V: Об’єм (Volume), Швидкість (Velocity), Різноманітність (Variety), Достовірність (Veracity) [8].

Методи боротьби з оголошеними проблемами представлені в наступних категоріях:

- Вибір методології навчання. Наприклад, Глибинне / Глибоке навчання (Deep Learning), Безперервне Навчання (Continuous / Lifelong Learning), Передача Навчання (Transfer Learning) і т.д.
- Маніпуляція з даними, наприклад, очищення даних, організація даних для навчання пакетом (batch) або потоком (stream), розширення кількості даних за допомогою часткової зміни існуючих даних, так звана аугментація даних, тощо.
- Техніки обробки самих даних, що передують етапу навчання, наприклад, вертикальне та горизонтальне масштабування системи навчання.
- Маніпуляція з алгоритмами навчання. Переважно основні зміни в алгоритмах навчання для підтримки обробки Великих Даних пов'язані з підтримкою паралелізації. Відомий приклад: технологія МепРід'юс (MapReduce) [10].

Організовуючи у впорядковану послідовність згадані вище методи, можна побудувати загальний алгоритм ефективного освоєння Великих Даних (рис. 1)

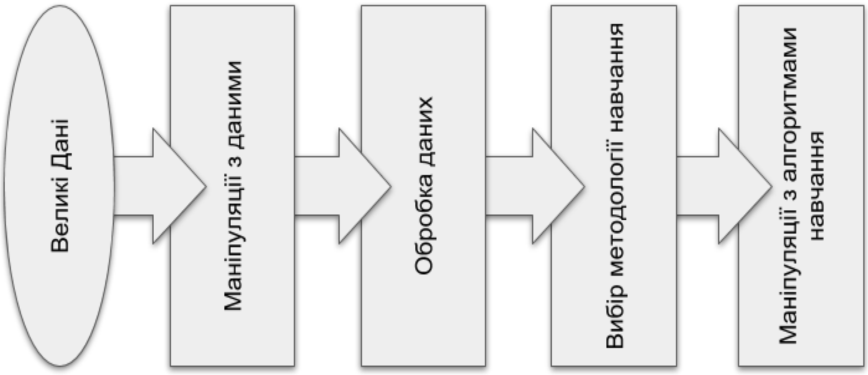


Рис. 1. Послідовність застосування методів обробки Великих Даних

Застосовуючи будь-які з вищеописаних методів окремо або використовуючи їх, як проілюстровано, у комбінації для усунення 4V проблематик, це – поточний спосіб освоїти переваги Великих Даних і принести якісно нову користь у ті прикладні області, де алгоритми машинного навчання міцно посіли своє місце.

Основні проблеми Машинного Навчання у контексті Великих Даних

Розглянемо детальніше кожен із чотирьох оголошених викликів для машинного навчання у середовищі Великих Даних і методи боротьби з ними.

Об'єм (Volume)

Напевно, найбільш репрезентативна характеристика Великих Даних це – їх об'єм. Об'єм даних на пристрої зберігання інформації, що його містить. У середовищі Машинного Навчання зростання обсягу розглядається як зростання кількості даних у двох напрямках:

- Горизонтальному – набір атрибутів, т.зв. ознак (features), які характеризують розглянуту інформацію.
- Вертикальному – кількість записів у наборі даних, тобто кількість прикладів навчання.

Зі збільшенням обсягу даних, як у вертикальному так і горизонтальному напрямках, зростає час, що витрачається на обробку даних, на відбір репрезентативних ознак, варіативність у налаштуванні гіперпараметрів системи навчання. Все це негативно позначається на продуктивності алгоритмів машинного навчання.

Логічною відповіддю на виклик, пов'язаний зі зростанням обсягу в горизонтальному і вертикальному напрямках, є зменшення розмірності даних і вибір найбільш показових примірників даних.

Простим, і в той же час поширеним, прикладом методу зниження розмірності є метод Аналізу Головних Компонентів (Principal Component Analysis – PCA), метод Аналізу Незалежних Компонент (Independent Component Analysis – ICA), Сингулярний розклад матриці (Singular Value Decomposition – SVD) та інші. Із сучасних підходів у завданні зниження розмірності значно досягли успіху такі інструменти як Автокодировщики (Autoencoders) [11].

Підходи з відбору репрезентативних ознак, т.зв. Інженерія ознак (feature engineering) [12] з розвитком методів Глибокого Навчання стають менш актуальними, тому що ці методи спроможні самостійно виділяти ознаки в даних під час навчання, переводячи дані в абстрактне уявлення і в такому вигляді вивчати їх, без явного ручного відбору – автоматично [13].

Що ж стосується методів відбору репрезентативних даних, то можна згадати випадковий, прогресивний, кластерний відбори даних [14]

Також з ростом обсягу даних, що надходять, можна розраховувати на допомогу від горизонтального масштабування системи навчання через розпаралелювання її обчислювальної інфраструктури на доступні вузли за допомогою технології МепРід'юс [10].

Різноманітність (Variety)

Під різноманітністю розуміють як різнорідність вхідних даних – різноманітність джерел, з яких з'являються дані, так і різнорідність самих типів даних. Також, до проблеми різноманітності мають відношення зашумленість і забрудненість даних. Основними методами боротьби з оголошеними проблемами, окрім очевидного очищення даних, використовуються певні переваги методологій навчання.

Для очищення даних і шумозаглушення в контексті Великих Даних користуються тими ж методами, що і для зниження розмірності простору ознак, згаданих у секції проблеми Обсягу.

Вибір методологій навчання для вирішення проблеми різноманітності даних лежить, переважно, серед таких методів: Передача Навчання, Глибинне Навчання, Безперервне Навчання.

Методологія Передачі Навчання корисна в контексті проблеми, що розглядається, через здатність витягти знання з однієї

або декількох схожих завдань і пе-редати їх до цільової задачі, минаючи повноцінний процес навчання [15].

Через особливості алгоритмів Глибинного Навчання автоматично вибирати репрезентативні ознаки в даних, вони розглядаються як ефективні методи при боротьбі з різнорідними даними, тому що здатні однаково добре відбирати ознаки з різнотипних даних і ігнорувати непоказові приклади у даних [13].

Алгоритми Безперервного Навчання мають на меті добувати і узагальнювати нову інформацію, спираючись на наявні знання, і будуть додатково згадані нижче.

Швидкість (Velocity)

У цьому випадку швидкість – це характеристика, з якою дані прибувають до місця свого зберігання і обробки, і яка дає відносну оцінку аналітикам: наскільки швидко вони прибувають. Очевидно, у цьому контексті важливою є і швидкість обробки цих даних. Часто буває так, що на обробку даних виділено обмежений фрагмент часу, після якого оброблений результат вже не такий важливий, або не важливий зовсім. Наприклад, котирування цін акцій, прогнози стихійних лих і т.д. Іншими словами, обробка нових даних, що надійшли, повинна бути своєчасною, тобто, коли нові дані надійшли, тоді ж, негайно, вони і повинні бути оброблені.

Більшість класичних алгоритмів машинного навчання розраховують на формування пакета даних і доступність всього набору даних для запуску процесу навчання. З огляду на згадані вище обставини, немає можливості чекати, коли цей пакет формується, адже цей процес недетермінований. Гарним прикладом вирішення цієї проблеми є алгоритми онлайн навчання, які не розраховують на формування пакета для запуску процедури навчання. Такі алгоритми використовують потік даних і безперервно модифікують свою модель [16].

На додаток до вищеназваних методів онлайн навчання, існують і схожі методи т.зв. Інкрементального Навчання (Incre-mental Learning), які, замість модифікації моделі алгоритму, з кожним новим прикладом накопичують навчальний пакет з потокових даних і запускають процедуру навчання на ньому.

У рамках підходів онлайн, Інкрементального Навчання, при оновленні моделі знань виникає проблема, т.зв. Дрифт концепції (Concept Drift). Коротко, під час донавчання екземпляри одного класу з потоку можуть зустрічатися набагато частіше, ніж екземпляри іншого класу, що буде впливати на рівномірність розподілу навчальних даних і обов'язково негативно позначиться на подальшому використанні алгоритму [17].

Існує ряд інструментів, як серед рішень для підприємств, так і рішень з відкритим кодом, що реалізують обробку даних потоками. У середовищі рішень для підприємств можна згадати StreamInsight від Microsoft Inc., InfoSphere Streams від IBM, у середовищі рішень з відкритим кодом варто згадати про рішення, що розвиваються у рамках некомерційної організації Apache Software Foundation, таких як Spark, Storm, Samza. Звичайно, це далеко не повний список доступних сервісів, але важливо розуміти, що рішення існують, як для дослідницьких, приватних цілей, так і для комерційних організацій різного масштабу.

У доповненні, як очевидний метод боротьби з даними, що швидко прибувають, служитиме збільшення потужності обчислювальної системи алгоритмів навчання, поширюючи кількість обчислювальних вузлів і застосовуючи модель розподілених обчислень МепРід'юс.

Достовірність (Veracity)

Раніше, коли даних для алгоритмів машинного навчання було значно менше, ніж доступно зараз, вони були якісно вивірені фахівцями так, що можна було їх вважати правдивими і з упевненістю відправляти до алгоритмів навчання. Зараз ситуація змінилася, зважаючи на труднощі, пов'язані з обсягом, швидкістю, різноманітністю Великих Даних, проблема

їх достовірності стала досить реальною, щоб про неї почати говорити [18]. Достовірність даних вказує наскільки ці дані точні, правдиві і що варто зробити, щоб усунути такі проблеми як зміщення, неузгодженість, дублювання і мінливість у них.

Як і у випадку з різнорідними даними, описаними в розділі Різноманітність та Об'єм, для вирішення проблем, пов'язаних із недостовірними даними, можна скористатися тими ж підходами, що і для вирішення проблем зашумленості, забрудненості, великої розмірності даних, а саме: очищення, шумозаглушення і зниження розмірності простору ознак даних.

Спираючись на особливість Глибинного Навчання, самостійно виділяти з даних показові ознаки, можна розраховувати, що ця особливість дозволить зосередити увагу навчальної системи на значущих, валідних прикладах з вибірки даних, а недостовірні приклади під час навчання будуть проігноровані.

Також, спираючись на тривалість процесу навчання, пов'язаного зі специфікою технік Безперервного Навчання, можна вважати, що недостовірні дані незначно вплинуть на загальну продуктивність процесу навчання.

У доповненні до оголошених труднощів, кажучи про достовірність даних, згадують також і проблеми, пов'язані з невпевненістю у даних і з т.зв. походженням даних [8].

Прогресивні підходи для боротьби з проблематикою Великих Даних

Консолідуючи всю вищеописану інформацію, представимо її у вигляді загальної схеми взаємодії (рис. 2).

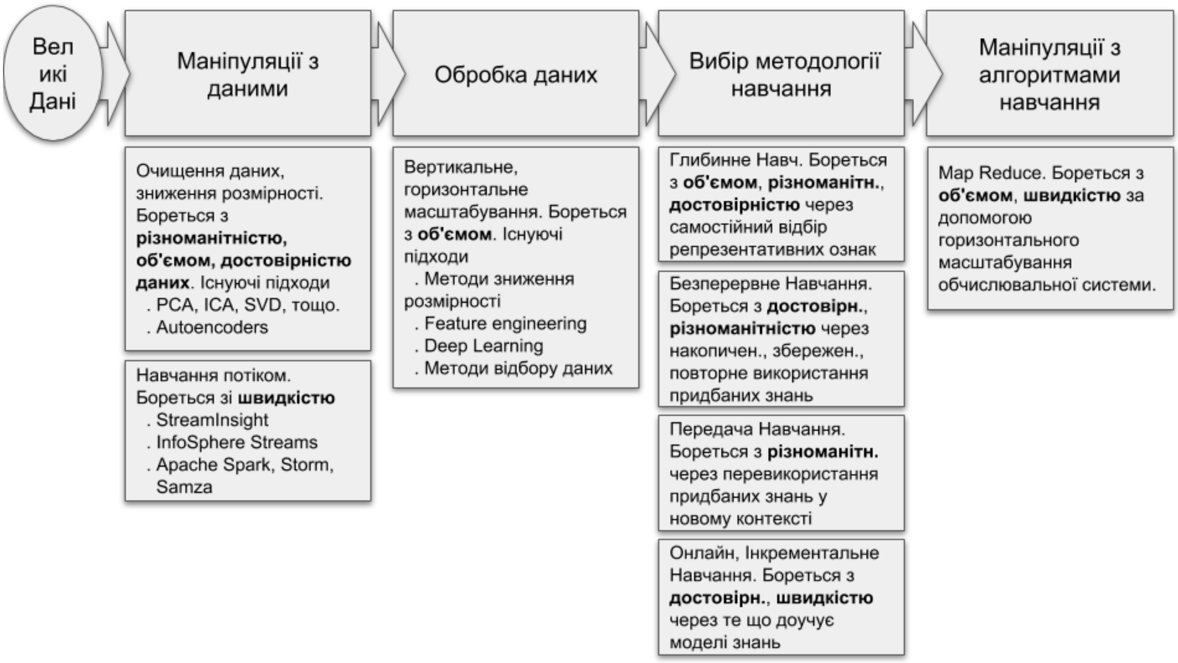


Рис. 2. Послідовність методів обробки Великих Даних з описом інструментів, технік і проблематик, яким вони адресуються

Як проілюстровано на рисунку 2, вибір методології навчання являє собою найповніший інструментарій для боротьби з оголошеними проблемами Великих Даних.

Окремо з вибору методології навчання хотілося б розглянути цілі і методи Безперервного Навчання. Як згадувалося раніше, основним призначенням методів цих підходів є утримання накопичених і виділення нових знань [19], чим окремо, у тій чи іншій мірі, займаються методи онлайн навчання і Передачі Навчання. Також підходи Безперервного Навчання

можуть приховувати в собі різноманітні алгоритми, зокрема алгоритми Глибинного Навчання, користуючись усіма їх перевагами. З цього боку Безперервне Навчання видається більш універсальним інструментом для розвитку інтелектуальних систем у контексті Великих Даних.

Якщо дивитися на методологію Безперервного Навчання як на один з підходів у Машинному Навчанні, то варто її диференціювати залежно від способу навчання, тобто Безперервне Навчання методами з та без вчителя та навчання з підкріпленням (Reinforcement Learning). Коротко наведені приклади по кожному з них.

В умовах навчання з учителем варто згадати т.зв. нейронну мережу, засновану на роз'ясненні – EBNN (Explanation-based neural networks). Така мережа спроможна передавати знання через декілька завдань, що навчаються, використовуючи доменні знання, отримані з попередніх завдань, і керуючи узагальненням знань при зустрічі з новими завданнями [20].

Говорячи про Безперервне Навчання в контексті навчання без вчителя, варто звернути увагу на роботу [21], де згадується багатошарова нейронна мережа зі спеціалізованої Машини Больцмана, навченої без учителя, і автоенкодери, здатні ефективно створювати ієрархію ознак, що уловлюють закономірності у вхідних даних так, як це відбувається у корі головного мозку людини в області, яка відповідає за функцію зору.

У разі навчання з підкріпленням хотілося б торкнутися роботи [22], що описує метод Безперервного Навчання у середовищі завдань навчання з підкріпленням, який спроможний на базі вже вивчених простих навичок виконувати набір більш складних.

У кожному зі згаданих способів Безперервне Навчання представлено окремо, але кожен зі способів стикається з однією і тією ж дилемою – баланс між стійкістю і гнучкістю у здобутті нових знань [23]. Дилема, що описує компроміс при навчанні в розподілених і паралельних системах. Для придбання нових знань необхідна пластичність системи навчання, щоб гарантувати інтеграцію нових знань поряд зі старими. Надмірна пластичність призводить до того, що раніше придбані знання просто забуваються системою і їх місце займають нові знання, це приклад т.зв. Проблеми Катастрофічного забування [24]. Для утримання старих знань необхідна достатня стійкість системи, але її надмірність заважає засвоювати нові знання.

Висновок і наступні кроки у дослідженні

У цій статті представлений відправний огляд технік Машинного Навчання для боротьби з проблематикою, викликаною Великими Даними. Перераховано ряд існуючих викликів у роботі з Великими Даними, методами традиційного Машинного Навчання. Дано обґрунтування, чому саме 4V проблеми є основними і поточними проблемами в освоєнні переваг Великих Даних. Представлений ряд методів для боротьби з кожною із оголошених проблем і обраний один із загальних і відповідних методів Машинного Навчання в даному контексті – Безперервне Навчання. У доповненні представлений короткий огляд існуючих проблем в обраній області Безперервного Навчання.

References

1. The Top 20 Valuable Facebook Statistics, (2018). Zephoria, *Digital Marketing*. Available from: https://zephoria.com/top-15-valuable-facebook-statistics/
2. Aslam, S. (2018). Twitter by the Numbers: Stats, Demographics & Fun Facts. Available from: https://www.omnicoreagency.com/twitter-statistics/
3. Mathison, R. (2018). 22+ Useful Instagram Statistics for Social Media Marketers; *Hootsuite*. Available from: https://blog.hootsuite.com/instagram-statistics/
4. Hernandez, D. How much Data will The Internet of Things (IoT) Generate by 2020? *Versa Technology*. Available from: https://www.versatek.com/blog/how-much-data-will-the-internet-of-things-iot-generate-by-2020/
5. Grolinger, K., Hayes, M., Higashino, W.A., L'heureux, A., Allison, D.S., Capretz, M.A.M. (2014). Challenges for MapReduce in big data *in Proc. IEEE World Congr. Services (SERVICES)*, pp. 182189.
6. Jagadish, H.V. et al. (2014). Big data and its technical challenges *in Commun. ACM*, vol. 57, no. 7, pp. 8694.

7. Chen, X.W., Lin, X. (2014). Big data deep learning: challenges and perspectives. *IEEE Access 2*, 514-525
8. L'heureux, A., Grolinger, K., Elyamany, H.F., Capretz, M.A.M., (2017). Machine Learning With Big Data: Challenges and Approaches, *IEEE Access(vol. 5)*, pp. 7776.
9. The Four V's of Big Data, IBM Big Data & Analytics Hub. Available from: http://www.ibmbigdatahub.com/infographic/four-vs-big-data
10. Grolinger, K., Hayes, M., Higashino, W.A., L'heureux, A., Allison, D.S., Capretz, M.A.M. (2014). 'Challenges for MapReduce in big data *IEEE World Congr. Services (SERVICES)*, pp. 182–189.
11. Hinton, G.E., Salakhutdinov, R.R. (2006). Reducing the Dimensionality of Data with Neural Networks *Science* – 2006-07-28. – Vol. 313, – P. 504–507. DOI:10.1126/science.1127647
12. Avrim, L. Blum, Langley, P. (1997). Selection of Relevant Features and Examples in Machine Learning, *Artificial Intelligence*, Volume 97, Issues 1–2, pp. 245-271
13. Bengio, Y., Courville, A., Vincent, P. (2013). 'Representation learning: A review and new perspectives *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 8, pp. 1798–1828.
14. Liu, H., Motoda, H., (2013). Instance Selection and Construction for Data Mining, *New York, NY, USA: Springer.* vol. 608.
15. Pan, S.J., Yang, Q. (2010). A Survey on Transfer Learning, *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359.
16. Shalev,-Shwartz S. (2012). Online Learning and Online Convex Optimization *Foundations and Trends® in Machine Learning*: Vol. 4: No. 2, pp. 107-194.
17. Gama, J., Zliobaite, I., Bifet, A., Pechenizkiy, M., Bouchachia, A. (2013). A Survey on Concept Drift Adaptation *ACM Computing Surveys*, Vol. 1, No. 1, Article 1.
18. Lukoianova, T., Rubin, V.L. (2014). Veracity Roadmap: Is Big Data Objective, Truthful and Credible? *Advances In Classification Research* Online, Conference Proceeding.
19. Daniel, L. Silver, Yang, Q., Li, L.H. (2013). Lifelong Machine Learning Systems: Beyond Learning Algorithms, *AAAI Spring Symposium: Lifelong Machine Learning*, Vol. 13 p. 05.
20. Thrun, S. (1996). Explanation-based Neural Network Learning: A Lifelong Learning Approach, *MA: Kluwer Academic Publishers* Boston.
21. Bengio, Y. Learning deep architectures for AI, *Foundations and Trends in Machine Learning* 2(1):1-127.
22. Ring, M.B. (1997). Child: A first step towards continual learning, *Machine Learning*, pp. 77-104.
23. Mermillod, M., Bugaiska, A., Bonin, P. (2013). The stability-plasticity dilemma: investigating the continuum from catastrophic forgetting to age-limited learning effects, *Frontier Psychology*.
24. Kirkpatrick, J., Pascanu, R., Rabinowitz, N., Veness, J., Desjardins, G., Rusu, A.A., Milan, K., Quan, J., Ramalho, T., Grabska-Barwinska, A., Hassabis, D., Clopath, C., Kumaran, D., Hadsell, R. (2017). Overcoming catastrophic forgetting in neural networks, *PNAS. 201611835*; published ahead of print March 14, 2017.

RESUME

V.M. Tereshchenko, A.D. Bugaiov

Machine Learning algorithms in Big Data context

The data isn't information but it carries information. In addition to the public information of the data that is clearly visible to a user there is hidden information, such as trends, user preferences, products and services satisfaction, etc., available only after a careful analysis of vast amounts of data. The amounts and sources of data grow each year with astounding speed, giving the opportunity to analyze and reveal hidden information.

One of a recognized successful method for obtaining meaningful information from a data is a Machine Learning (ML), however there are existed enough technical challenges and difficulties.

As recognized difficulties of ML in the Big Data context are the 4V's problems: Volume, Velocity, Variety, Veracity.

The amount of data on the occupied storage device is a volume characteristic. As an amount of data grow the time spending on data processing grow also and negatively affect on performance of ML algorithms. The variety is the diversity of sources from the data are appeared, the heterogeneity of the data types and noisy and dirty data.

The velocity is the speed with which the data arrives to a place of its storage and processing and gives a relative estimation how quickly the data arrive. In this context the speed of processing data is also important as after some time if the data have not been processed the result is no important anymore.

In the view of the difficulties associated with the volume, velocity, variety, the problem of data reliability has become real for involving one more problematic like veracity.

Regarding all the problems above as one of a promising way to address them are Lifelong Learning algorithms. The Lifelong Learning algorithms are presented in every of ML methods, such as Supervised, Unsupervised and Reinforcement methods, but every of them is encountered the same stability-plasticity dilemma. The algorithms are proposed to consider as a tool set is able to solve more than one from the announced problems in context of Big Data

Надійшла до редакції 22.10.2018